DOCUMENT RESUME

ED 147 361 TM 006 815

| | |
|---|---|
| TITLE | Guidelines and Cautions for Considering Criterion-Referenced Testing. |
| INSTITUTION | National Education Association, Washington, D.C. |
| PUB DATE | Aug 75 |
| NOTE | 21p.; Also available in ED 146 233 |
| AVAILABLE FROM | National Education Association, 1201 16th Street, N.W., Washington, D.C. 20036 (Price not available) |
| EDRS PRICE | MF-$0.83 Plus Postage. HC Not Available from EDRS. |
| DESCRIPTORS | *Achievement Tests; *Criterion Referenced Tests; *Diagnostic Tests; *Educational Objectives; *Guidelines; Norm Referenced Tests; Standardized Tests; Standards; *Teacher Role; Test Bias; Test Construction |
| IDENTIFIERS | Domain Referenced Tests; *Objective Referenced Tests |

ABSTRACT

The opinions presented reflect those of the National Education Association Task Force on Testing: that norm referenced tests are often misused, and that problems are also associated with criterion referenced tests. A list of warnings for teachers considering the use of criterion referenced tests includes: (1) common deficiencies in testing need to be communicated both to the profession and to the public, (2) teachers should have an extensive role, from the beginning, in determining objectives, (3) the claims of criterion, objective, and domain referenced tests should be viewed with some skepticism, but with an open mind, (4) teachers should obtain information on field testing, reliability, and validity of the tests they use, (5) teachers should vigorously resist the misuse of all kinds of tests, and (6) teachers should not allow themselves to be evaluated according to the results of any test. (Author/CTM)

**********
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
**********

TM

ED147361

# GUIDELINES AND CAUTIONS FOR CONSIDERING

# CRITERION-REFERENCED TESTING

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
National Education Association
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Published by the

NATIONAL EDUCATION ASSOCIATION
1201 - 16th Street, N.W., Washington, D.C. 20036

August 1975

TM006 815

## PREFACE

When the NEA Task Force on Testing began to conclude that the great preponderance of standardized tests likely do more harm than good, they initiated an exploration of alternatives. In doing so, they recognized the potential of criterion-referenced testing for alleviating some of the destructive effects of standardized measures and engaged in an examination of the concept.

The Task Force's conclusions about the usefulness of criterion-referenced tests (CRT's) resulted in mixed reviews. While they were aware that teachers have long used CRT's in some form (the Friday quiz based on the instructional objectives for the week is one example), they hoped to identify more refined adaptations. Increased psychometric sophistication was indeed found, but frequently it was without correction of the major deficiencies in standardized tests. As a result of their deliberations and findings, the Task Force gave direction to the development of the guidelines and cautions that follow. The 15 caveats present a strong point of view on criterion-referenced testing, one to which teacher associations will want to give serious attention where criterion referenced tests are proposed or already in use.

For those who need to brush up a bit on some commonly used definitions related to tests, a glossary of measurement terms is appended.

Bernard H. McKenna
Professional Associate
NEA Instruction and Professional Development

GUIDELINES AND CAUTIONS

FOR CONSIDERING CRITERION-REFERENCED TESTING

Standardized achievement tests used in most schools today are known as "norm-referenced" tests. They are constructed in such a way as to maximize differences among students so that one can be compared to another. This is done by providing for maximum discrimination between high and low scores. The purpose is to rank a student among his peers. Hence, scores are reported in such terms as "John Jones is in the ninety-fifth percentile on verbal reasoning." While norm-referenced tests are useful for sorting people into categories (to the dismay of many), they are not useful for improving educational programs.

Recently a new concept has been promoted among test makers and the educational public called criterion-referenced testing, also termed objective-referenced testing. At least three factors have contributed to emergence of the concept: First is a strong and rising dissatisfaction with tests in general. Second is the inadequacy of traditional tests for diagnostic and instructional purposes. Third, there is some clamor for evaluating instruction and teachers as part of the accountability movement. Although criterion or objective-referenced tests may have potential for diagnosing learning problems and improving instruction, they are not useful for evaluating teachers. (Test scores depend largely on variables in a student's background rather than on what he or she is taught in the classroom. Even so, a few years ago a bill was introduced in the Kansas legislature to cut off funds to districts whose children did not score above the national average on such tests. Fortunately the bill did not pass.)

Criterion-referenced tests, instead of comparing one child to another, presumably measure the child's performance against a specified criterion (or

objective). Thus all children might be able to achieve the criterion and eventually score 100 percent on the tests. The criterion-referenced test, in concept, is much like the kind of test the teacher gives in the classroom on Friday to evaluate learning of specific objectives taught earlier in the week.

Conceivably the external criterion toward which the test is directed could be a number of things. For example, one could have a criterion-referenced test for measuring the skills of a bricklayer (no doubt a nonverbal test) -- Can he lay bricks? Can he mix mortar? -- without reference to how others do. The higher an individual scored on the test, the closer he would be to acquiring a bricklayer's skills, regardless of how many other people had the same skills.

Test makers, however, have shown little inclination to develop tests directed toward such criteria. Establishing a sequence of skills and validating them is a laborious, difficult, multiyear task at best. Staying with the example of the bricklayer, they would have to conduct studies to show that good bricklayers score high on the test; that is, they would have to evaluate the test. Test makers instead have resorted to a conception of criterion-referenced tests as those which yield measurements "directly interpretable in terms of specified performance standards" (Glaser and Witko, 1971). In practice, this means that the criterion toward which the test is directed is usually a prespecified objective (an objective stated in advance, e.g., "A bricklayer must be able to mix mortar").

Thus "criterion-referenced" usually means in practice "objective-referenced." In fact, those who have most strongly propagated criterion-referenced testing are frequently the same persons who have propagated behavioral objectives. In typical procedure, objectives are established and test items are written to measure those objectives. Test results can be reported in terms of what specific objectives each individual student was able to achieve, which presumably is useful

for instructional purposes. In this way, it is argued, tests can be tailored to specific objectives the way a teacher tailors test questions on what he or she has taught.

The distinction between criterion-referenced and norm-referenced tests is quite blurred. Most test makers use similar procedures to construct items for both types, or use the same item, and employ test statistics for norm-referenced items in selecting items for criterion-referenced tests. There are no clearly defined and commonly agreed upon procedures for constructing criterion-referenced tests, and many of them are in fact norm-referenced tests in disguise. The distinction becomes a matter of emphasis rather than being clear-cut.

Womer (1973) defines a criterion-referenced test as --

> ...one which is designed to provide information about attainment of a specific objective (criterion), which emphasizes direct measurement through the use of differing formats, which may use items at varying difficulty levels, which must have content validity, which must minimize guessing, and which is particularly useful for instructional and evaluative purposes.

Womer's "differing formats" term indicates he is keen on test items which call for responses other than multiple-choice. Many criterion-referenced tests continue to be made up mainly of multiple-choice items.

A main advantage claimed for criterion-referenced tests is their utility for improving educational programs. In view of the confusion among test makers themselves about the concept, construction, and utility of the tests, some caveats are in order for those considering the use of criterion-referenced or objective-referenced tests.

1. Common deficiencies in testing need to be communicated both to the profession and to the public. Neither criterion-referenced tests (CRT's) nor objective-referenced tests (ORT's) eliminate the most common deficiencies of tests in general.

CRT's and ORT's for the most part still measure simple tasks at the expense of relearning abilities and higher-level thought processes (Stake, 1973). Complex performances are so difficult to measure that test items reflect only the simpler tasks. Such things as Binet's categories of mental imagery, imagination, aesthetic appreciation, and moral sensibility are almost totally unmeasured.

2. Teachers should examine carefully the derivation of the objectives for ORT's.

ORT's can be no better than the objectives on which they are based. Unfortunately, the methods for deriving objectives are often ill-considered, hasty, and grossly inadequate. There is an inclination among test makers to slide over the problems of deriving objectives in order to get to item construction, a task with which they are more familiar. Yet appropriate objectives are just as important and just as difficult to arrive at as are test items.

There are at least four ways to choose objectives (Klein and Kosecoff, 1973). First, choosing by expert judgment means that a small group of subject matter experts decides which objectives should be measured for a given field. This was essentially the origin of National Assessment tests. While few persons would deny the relevance of the judgments of subject matter experts, few would contend that such judgments faithfully or completely represent what should be taught. By no means do they fully represent the judgments of teachers, parents, students, and others vitally concerned.

A second way of choosing objectives is by consensus judgment, which requires that various groups -- teachers, administrators, parents, school board, etc. -- decide what objectives are most important. (In this paper, "objectives" refers to specific student learning outcomes.) Unfortunately, the immense problems of

such prioritizing have been slighted. Frequently decision-making groups respond only to those objectives that are presented to them by a single group (e.g., school administrators) or a limited number of groups. Correcting for important objectives that have been omitted is not taken into account. If critical objectives do not emerge from the objective-generating process they are ordinarily lost forever. For example, there is likely to emerge a high preponderance of content-bound objectives that are easily measurable. More subtle learnings are neglected. Attending to the objectives that are easily identifiable severely limits the range of decision-makers' thinking and results in determining (and limiting) the curriculum.

The rating of priority statements themselves is severely dependent upon how abstractly the objectives are specified (how global they are), the types of criteria on which the objectives are rated (Are they rated in "importance," how much money will be spent on them, how much time and effort will be spent, and the nature of the groups doing the rating?) (Stake and Gooler, 1973; House, 1973). Test makers have had little experience polling the opinions of non-professional groups, so surveys for the purpose of developing or rating the importance of objectives are likely to be highly class-biased. Actually, such surveys are seldom done. Objectives generation and measurement are likely to be treated in the most cavalier fashion. Test-developers who would never think of including an item without field testing it sometimes accept and discard objectives with abandon. A common procedure is to have the objectives reviewed by a small group of citizens and educators and claim that the objectives have been approved by the public. Those citizens involved are too frequently upper-middle-class and the educators so selected that they are not broadly representative.

A third way of deriving objectives is through curriculum analysis. One can inspect materials such as textbooks or courses of study to determine what is being taught and then write objectives and test items based on such content. Much of

the impetus for CRT's came from curriculum developers like those who pioneered Individually Prescripted Instruction (IPI) as part of their efforts to develop tests that measure exactly what the materials teach. This procedure also has its limitations in that it is likely to emphasize only content-related objectives.

Fourth, objectives can be chosen by in-depth analysis or those instructional areas which one wishes to test. One tries to determine the contents and behaviors in an area of instruction and to associate objectives and test items with contents and behaviors. In other words, by task analysis the instruction is broken into discrete learnings. The most ambitious efforts along this line have resulted in instruments called "domain-referenced tests" (Hively, 1973; Baker, 1973; Millman, 1973).

Domain-referenced testing (DRT) attempts to define "domains" of behavior -- categories of behavior one might test and teach for -- and to represent these domains by an extensive pool of test items which measure human performance in a particular domain or domains. In one sense, domain-referenced tests appear to be an attempt to escape the triviality and absurdity of much of the behavioral objectives movement. If one must delineate a highly specific objective for each aspect of student behavior, one might generate thousands of such objectives. In one project an attempt to define a complete set of objectives for the high school was given up after 20,000 objectives had been written. A complete delineation becomes an absurdity and most such lists become trivial.

Domain-referenced testing aims at overcoming these problems by defining important categories of content and behavior so that only objectives representing particular domains become important. Other objectives are merely subsets or exampl The instructional benefits of such a scheme promise to be large since one could practice on other objectives and test items from the domain to learn the behavior. One could always construct another test from the innumerable objectives and test items representing that domain.

DRT's exist more in promise than in practice. No doubt the task analysts will confront the same formidable conceptual problems as psychologists who have tried to categorize mental behavior and curriculum developers who have tried to define the "structure" of their subject. Even the most sophisticated schemes of human mental abilities, such as Bloom's Taxonomy, tend to falter when subjected to empirical examination. Human mental processes defy categorization which suggests emphasis on the long-debated principle of teaching to the whole child rather than to specific skills.

3. Teachers should have an extensive role, from the beginning, in deriving objectives and should beware of co-optation.

Most teacher (and public) involvement in developing objectives has been cursory at best -- more for the purpose of legitimizing the objectives than for determining or implementing them. For example, objective-referenced tests were developed for the state assessment program in Michigan and employed on a mandatory basis at selected grade levels. For the selected grades, subject specialists from the state education agency set up a small committee of educators, including four teachers, to select and review objectives. The committees developed goals which were later reviewed by subject matter associations. Then several one-day large group meetings were held around the state to give people a chance to respond.

Despite this effort to involve them, many of the teachers and administrators who participated in the group meetings felt that they had not had adequate input on the objectives (House, Rivers, and Stufflebeam, 1974). They were presented with a list of objectives and asked to respond after a cursory review. Most teachers in the state never saw or heard of the objectives. In spite of promises that the objectives were only for experimental purposes, the state agency developed tests based on them and administered them the following year, claiming educator endorsement.

4. Which objectives are selected and retained for testing is critical for ORT's. Teachers should be intimately involved from the beginning in selecting objectives.

Selection of final objectives for testing is as important as generating them, and teachers are frequently provided only cursory participation in this activity also. In the Michigan assessment program over four hundred objectives were generated for fourth-grade mathematics, yet only thirty-five were selected for testing. The limiting factor was the amount of time required for testing each objective (it was deemed advisable not to exceed five hours of testing time). Which objectives were excluded? Why? If only the most important objectives were included, how was "importance" determined? What would be the instructional effect over time of excluding the other several hundred objectives? In most cases of objective development the objectives are rewritten and screened by state education agency officials, select citizens' groups, and test makers. For example, in Illinois, goals derived from public hearings were selected and extensively rewritten by several groups before being presented as public goals.

5. The ways in which test items are constructed should be examined. When possible, teachers should employ their own test experts to help them assess the procedures.

The usual number of items to measure one objective seems to vary from three to five. Good results have been obtained with five. Since even the most specific objective can be measured by thousands of test items, selection is important. Sophisticated test makers use a systematic sampling plan that produces items for subcategories of the objectives.

Of at least equal importance is the type of response the item calls for. Traditional tests use multiple-choice answers because they are easy to machine-score. However, if the purpose of the test is to describe and diagnose classroom

learning and provide usable information to the teacher, multiple-choice answers may be much less desirable. The degree to which a test is a faithful sample of learning behavior is more important in an objective-referenced test than in one which merely strives to differentiate among students.

A group of items constructed by teachers is likely to be more relevant to the instruction of those particular teachers. Items written by measurement experts from a matrix of content and behavior are likely to be technically better but less relevant.

6. CRT's and ORT's should be thoroughly field tested. Teachers should refuse to use tests that have not been thoroughly field tested.

While this may seem a rather obvious caveat, the fact is that many objective-referenced tests have not been extensively tried out. Even where tried out, frequently only a handful of students are involved. Tests with so little field testing should be resolutely avoided. The test developer should be required to present details of the field test. If he can't he probably hasn't conducted one, an all too common occurrence.

7. Test developers should present evidence of the test's reliability. Teachers should not use tests for which evidence on reliability is unavailable.

For an ORT, each set of items used to measure an objective might be considered a test in itself. These should be reliable measures in and of themselves. The usual reliable determinants are test statistics which are measures of internal consistency developed for traditional norm-referenced tests. They are based on variations in individual test scores -- item difficulty and the differences between the top scorers as opposed to bottom scorers, for example. The reliability will be highest when about half of the students get an item right and half get it wrong -- a norm-referenced concept maximizing discrimination among test takers.

Using these traditional techniques causes the tests to discriminate in the same way as do items in standardized tests. Unfortunately, the ORT developers have not been able to solve this problem. The alternative is to have no evidence of reliability, which to many is even more unacceptable. Perhaps the best policy is to insist on some measures of reliability, ones for which the test developers supply a public rationale which can be assessed.

8. The test makers should present evidence of the validity of the tests. Teachers should inspect the validation procedures carefully.

Validity -- which depends upon the ability to answer the question, "Does the test measure what it is supposed to?" -- presents another difficult problem for the maker of criterion-referenced tests. For traditional norm-referenced tests, validity is often established by how well the test predicts concurrent academic grades. But this makes little sense for CRT's. Test developers are usually left trying to make logical assessments of "content validity" based on how the tests were developed.

If the test is objective-referenced, one can assess whether test items adequately measure the objectives and whether the objectives themselves are valid for what the test is trying to measure.

If the test purports to measure the effects of classroom instruction, then the objectives must be the ones taught and the test items must be sensitive to instruction. The Michigan assessment program tried a "sensitivity index" to determine if correctly responding to an item was dependent on instruction. The index didn't work in this situation. A highly specific objective might be valid for one class but not for another, and a test which presumes to be valid for assessing instruction in a whole state has the problem of demonstrating that its items and objectives were constructed in such a way as to be appropriate statewide -- not an easy task. The whole problem of validity is an unresolved one, but the burden of proof should fall on the test maker, not the buyer.

No matter what the derivation of the test or what it is called, unless it covers what a particular teacher has taught it cannot be a valid measure for that teaching situation; it is a measure of someone else's objectives. On the other hand, if the test is a measure of objectives which the teacher developed but which he is willing to accept as indicative of his instruction, then the objectives are valid for that teaching situation.

9. "Minimal competency" or "mastery" cut-off points for students should be viewed with some suspicion. Teachers should question arbitrary standards and substitute their own.

Item difficulty on tests can be manipulated easily by test makers. Whether a student scores 30 percent or 88 percent can be built into the test itself and just as easily changed by assigning arbitrary values to test items. Since there is no objective means by which tests can establish a level of satisfactory "competency," the setting of such standards is extremely arbitrary. What is minimal competency in reading? When has one "mastered" reading? On the other hand, one may be willing to accept the opinions of certain groups as standards if they are clearly recognized as group opinion and subject to all the deficiencies that implies.

Nonetheless, many CRT developers continue to build highly arbitrary standards into their tests. For example, the Michigan assessment is based on a minimal skill concept that declares a student must achieve 75 percent of the minimal objectives. In the first year of implementation some of the districts where the highest academic achievement might be expected were able to achieve only 30 percent of some objectives. The 75 percent cut-off was evidently without justification.

10. Many objective-referenced tests are really norm-referenced tests in disguise. No teacher should voluntarily administer a test that he does not understand.

If one constructs objectives such as "reading a newspaper at a fourth-grade level," the norm is obviously built in. If one then selects test items using traditional test statistics, like item difficulty, and uses items from norm-referenced tests, the result is a test that discriminates among students but has the appearance of being referenced to skills rather than students. It becomes a norm-referenced test that looks like a criterion-referenced test. (Some test experts claim that it is impossible to construct anything other than a norm-referenced test.) It is also possible to use ORT results in a norm-referenced manner if one counts how many objectives each student learned and then makes comparisons among students.

11. The public and the profession should be made aware that CR or ORT's are not panaceas. Test bias problems remain the same with CR or ORT's as with norm-referenced tests.

Lower-socioeconomic groups will score as low on criterion or objective-referenced tests as they do on norm-referenced tests. Basic factors such as malnutrition and lack of motivation toward school and test taking are untouched by change from one type to another. What CRT's might offer some students is a reprieve from being told they are inferior. (In some districts test scores are attached to the report cards or even reported in the newspapers.) Since self-confidence seems to be critical in schooling, lack of stigmatization could be an important advantage. Another advantage might be to spell out in greater detail where certain educational weaknesses of students lie. Actually, CRT developers have done little that might result in preventing racial class, school building, or neighborhood bias in their tests.

12. CRT's could cost more than traditional tests, depending on the thoroughness of development. The costs of tests versus their utility should be carefully considered.

Traditional norm-referenced tests already exist and do not need to be developed, so if CRT superiority can't be positively demonstrated, the question should be raised, "Why go to the extra time and expense?" Also, because of their greater specificity, consider that CRT's might be valid for only a small domain of behavior at a given point in time (there could be large rewards in this, of course, in promoting learning). Many more tests would have to be developed rather than a few general ones. The procedure of developing and validating objectives and test items is a long, difficult, and costly procedure when properly done.

There are two ways of reducing costs. One is based on the assumption that there are certain basic and necessary skills and stages of learning independent of the local setting and that one need develop only one test for basic reading skills and sell it to everyone. This is the assumption of the test makers -- but it is a questionable one. Learning often seems to be highly context-dependent. Children learn in different ways in different settings. The inability of educational research to come up with guaranteed teaching techniques and the inability of psychology to demonstrate transfer of training indicates this is so.

Another way of reducing costs would be to have local groups of teachers develop their own CRT's as they now do for their classrooms. But there is the question of whether the amount of time required would be profitably spent in test construction. (See chapter 11, "Cooperative Development of Evaluation Systems for Student Learning," in Bloom, Hastings, and Madaus, 1971.)

13. Teachers should not be evaluated on CRT's and ORT's any more than on norm-referenced tests. Teachers should not allow themselves to be evaluated on the basis of ANY tests.

Tests are not good measures of what is taught in school. Although objective-referenced tests purport to be better measures of learning, they cannot be considered good measures of teaching. An obvious deficiency is that the tests measure only cognitive aspects of the classroom. In addition, the teacher does not have control over many of the variables that affect test scores. Evaluating teachers is a use that should not be claimed for ORT's. The evaluation of teaching should be based on observation, self-evaluation, student ratings, interviews, and many other types of data.

14. A main advantage of CRT's or ORT's seems to be in the reporting of results, that is, avoiding blanket categorizations of children by test scores and providing more useful instructional information. Subtests should be used only as diagnostic instruments.

Instead of a composite score with which the teacher can do little but type the child, in criterion or objective-referenced testing the teacher is presented with specific objectives the student can or cannot accomplish. The avoidance of a single score categorizing the child is a major benefit. Presumably the teacher also will be better able to make use of the detailed objectives for improving instruction and learning.

It should be noted, however, that there is little evidence that a teacher can do a better job working with specific objectives than working without them. Whether to use specific objectives should remain a matter of style and judgment for the individual teacher. Stake (1973) has indicated that there are significant costs in using behavioral objectives, including the possibility that the teacher will teach only what is easy to measure. In Michigan, most teachers did not find

the ORT's valuable for instructional purposes (House, Rivers, and Stufflebeam, 1974). The instructional benefits are also reduced by the limited number of objectives to which one can teach and for which one can reasonably test.

15. While worthy of consideration, the claims of criterion, objective, and domain-referenced tests should be viewed with some skepticism but with an open mind. Teachers should vigorously resist the misuse of all kinds of tests.

In some ways CRT's can be viewed as a response by the testing establishment to avoid some of the criticisms of tests. Such was the motivation in Michigan. CRT's and ORT's still embody most of the deficiencies of tests in general and are not useful for evaluating teachers in accountability schemes. The tests are also difficult to construct and are subject to much conceptual confusion, even though they do offer the potential of being more useful for instruction.

An important benefit of CR versus norm-referenced tests is that with CRT's the test taker is not stigmatized by a global score supposedly representing his/her ability. This is a great advantage. The best use of tests is in raising questions in the teacher's mind about individual students who achieve unusual scores. The tests themselves may be in error, or the teacher's preconception may be. In any case, following up on seeming discrepancies is the job of the professional. Tests should be used to raise questions, not to resolve them.

## REFERENCES

Baker, Eva L. "Beyond Objectives: Domain-Referenced Tests for Evaluation and Instructional Improvement." Educational Technology, 1973.

Bloom, Benjamin J.; Hastings, J. Thomas; and Madaus, George F. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill Book Co., 1971.

Glaser, Robert, and Witko, Anthony J. "Measurement in Learning and Instruction." Educational Measurement. (Edited by Robert L. Thorndike.) Washington, D.C.: American Council on Education, 1971. pp. 625-70.

Hively, Wells. "Domain-Referenced Testing." Educational Technology, 1973.

House, Ernest R. "Validating a Goal-Priority Instrument." Paper presented at the annual meeting of American Educational Research Association, New Orleans, February 25-March 1, 1973.

House, Ernest R.; Rivers, Wendell; and Stufflebeam, Dan. An Assessment of the Michigan Accountability System. Michigan Education Association and National Education Association, March 1974.

Klein, Stephen P., and Kosecoff, Jacqueline. Issues and Procedures in the Development of Criterion-Referenced Tests. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement, and Evaluation, September 1973.

Millman, Jason. "How To Make Assessment Plans for Domain-Referenced Tests." Educational Technology, 1973.

Popham, W. James, and Husek, R. R. "Implications of Criterion-Referenced Measurement." Journal of Educational Measurement, 1969.

Stake, Robert E. "Measuring What Learners Learn." School Evaluation. (Edited by Ernest R. House.) Berkeley, Calif.: McCutchan Publishing Corp., 1973.

Stake, Robert E., and Gooler, Dennis. "Measuring Goal Priorities." School Evaluation. (Edited by Ernest R. House.) Berkeley, Calif.: McCutchan Publishing Corp., 1973.

Womer, Frank B. "What is Criterion-Referenced Measurement?" IRA Committee on the Evaluation of Reading Tests.

# GLOSSARY OF MEASUREMENT TERMS*

## ACHIEVEMENT TEST

A test that measures the amount learned by a student, usually in academic subject matter or basic skills.

## APTITUDE TEST

A test consisting of items selected and standardized so that the test yields a score that can be used in predicting a person's future performance on tasks not evidently similar to those in the test. Aptitude tests may or may not differ in content from achievement tests, but they do differ in purpose. Aptitude tests consist of items that predict future learning of performance; achievement tests consist of items that sample the adequacy of past learning.

## CRITERION

A standard or judgment used as a basis for quantitative and qualitative comparison; that variable to which a test is compared to constitute a measure of the test's validity. For example, grade-point average and attainment of curricular objectives are often used as criteria for judging the validity of an academic aptitude test.

## CRITERION-REFERENCED TEST

A test in which every item is directly identified with an explicitly stated educational behavioral objective. The test is designed to determine which of these objectives have been mastered by the examinee.

## GRADE NORM

The average test score obtained by students classified at a given grade placement.

## LOCAL NORMS

Norms that have been obtained from data collected in a limited locale, such as a school system, county or state. They may be used instead of national norms to evaluate student performance.

## MULTIPLE-CHOICE ITEM

A test question consisting of a stem in the form of a direct question or incomplete statement and two or more answers, called alternatives or response choices. The examinee's task is to choose from among the alternatives provided the best answer to the question posed in the stem.

---

* Excerpts from the revised edition of A Glossary of Measurement Terms: A Basic Vocabulary for Evaluation and Testing, published by CTB/McGraw-Hill, Del Monte Research Park, Monterey, California 93940. Copyright © 1973. Reprinted by permission of the publisher.

NONVERBAL TEST

A test in which the items consist of symbols, figures, numbers, or pictures, but not words.

PERFORMANCE TEST

A test that requires the use and manipulation of physical objects and the application of physical and manual skills. Shorthand or typing tests, in which the response called for is similar to the behavior about which information is desired, exemplify work-sample tests, which are a type of performance test.

RANDOM SAMPLE

A sample drawn in such a way that every member of the population has an equal chance of being included, thus eliminating selection bias. A random sample is "representative" of its total population.

RELIABILITY

The consistency of test scores obtained by the same individuals on different occasions or with different sets of equivalent items; accuracy of scores. Several types of reliability coefficients should be distinguished.

Coefficient of internal consistency is a measure based on internal analysis of data obtained on a single trial of a test (Kuder-Richardson formulas and the split-half method using the Spearman-Brown formula).

Coefficient of equivalence or alternate forms reliability refers to a correlation between scores from two forms of a test given at approximately the same time.

Coefficient of stability or test-retest reliability refers to a correlation between test and retest with some period of time intervening. The test-retest situation may be with two forms of the same test.

STANDARDIZED TEST

A test constructed of items that are appropriate in difficulty and discriminating power for the intended examinees and that fit the preplanned table of content specification. The test is administered in accordance with explicit directions for uniform administration and is used with a manual that contains reliable norms for the defined reference groups.

VALIDITY

The ability of a test to measure what it purports to measure. Many methods are used to establish validity, depending on the test's purpose.